Multi-Pitch Estimation using Harmonic MUSIC

Christensen, Mads G.; Jakobsson, Andreas; Jensen, Søren Holdt

Published in:
Rec. Asilomar Conference on Signals, Systems, and Computers

Publication date:
2006

Document Version
Accepted author manuscript, peer reviewed version

MULTI-PITCH ESTIMATION USING HARMONIC MUSIC

Mads Græsbøll Christensen[*†], *Andreas Jakobsson*[‡], *and Søren Holdt Jensen*[†]

[†] Dept. of Electronic Systems
Aalborg University, Denmark
{mgc,shj}@kom.aau.dk

[‡] Dept. of Electrical Engineering
Karlstad University, Sweden
andreas.jakobsson@ieee.org

ABSTRACT

In this paper, we present a method for estimation of the fundamental frequencies, or pitches, of several periodic sources. This difficult estimation problem occurs, for example, in speech and audio processing whenever multiple speakers or tones are present. The presented method is an extension of a recently proposed method for estimation of the fundamental frequency of a set of harmonically related sinusoids based on the MUSIC criterion. The estimator takes into account that the number of harmonics and sources may vary. The performance of the proposed method is evaluated in Monte Carlo simulations using synthetic signals under various conditions.

1. INTRODUCTION

The task of estimating the fundamental frequency of a set of harmonically related sinusoids is a classical problem in speech and audio processing. Most research in the literature on this topic is devoted to the estimation of the parameters of a single set of harmonics, e.g. [1, 2, 3]. Multiple sets of harmonics occur when multiple speakers are speaking at the same time or when multiple notes are played simultaneously in music. For most of these methods, the extension to multiple sets of harmonics is not trivial. It has recently been shown that high-resolution fundamental frequency and order estimates can be obtained using subspace-based methods [4, 5], and, in this paper, we extend the method proposed in [4, 5], named harmonic MUSIC (HMUSIC), to multiple sets of harmonics. For other examples of multi-pitch estimators see, for example, [6, 7, 8]. Consider a signal consisting of several, say K, sets of harmonics (hereafter referred to as sources) with the fundamental frequencies ω_k, for $k = 1, \ldots, K$, that is corrupted by an additive white complex circularly symmetric Gaussian noise, $w(n)$, for $n = 0, \ldots, N-1$, i.e.,

$$x(n) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} A_{k,l} e^{j(\omega_k l n + \phi_{k,l})} + w(n), \qquad (1)$$

where $A_{k,l} > 0$ and $\phi_{k,l}$ are the amplitude and the phase of the l'th harmonic of the k'th source, respectively. The task at hand is then to estimate the fundamental frequencies $\{\omega_k\}$, or, equivalently, the pitches, from a set of N measured samples, $x(n)$. As a by-product of these estimates, also the set of orders, $\{L_k\}$, and the number of sources, K, are also estimated. The importance of estimating the order of the individual sources is twofold. Firstly, it is important since an incorrect order estimate may result in ambiguities in the cost function that may lead to gross errors in the fundamental frequency estimates. Secondly, its importance in finding high-resolution estimates can be observed from the asymptotic Cramér-Rao lower bound (CRLB) for the fundamental frequency, which for a single source, say source k, and a high number of samples, i.e., $N \gg 1$, can be shown to be [5]

$$CRLB_k = \frac{6\sigma^2}{N^3 \sum_{l=1}^{L_k} A_{k,l}^2 l^2}. \qquad (2)$$

As can be seen, the squared amplitudes of the individual harmonics are weighted by l^2, meaning that higher harmonics are in fact very important in obtaining an accurate fundamental frequency estimate. The expression in (2) is obtained under the assumption that the sources are independent and have well-separated fundamental frequencies. For a low number of samples, however, the exact CRLB for the fundamental frequencies will depend on the parameters of the other sources as well and on the fundamental frequency.

The paper is organized as follows. In Section 2, we present the fundamentals of the covariance matrix of the considered data model. Then, in Section 3, we present the proposed estimator and outline its efficient implementation. In Section 4, we evaluate the estimator and provide some illustrative examples of its properties. Finally, Section 5 concludes on our work.

2. FUNDAMENTALS

The proposed method is based on the MUSIC orthogonality principle [9, 10]. It is based on a partitioning of the eigenvalue decomposition (EVD) of the covariance matrix into signal and noise subspaces. We now briefly summarize the fundamentals of the covariance matrix model and the MUSIC orthogonality principle. We start out by defining $\tilde{\mathbf{x}}(n)$ as a signal vector containing M samples of the observed signal, i.e.,

$$\tilde{\mathbf{x}}(n) = \left[\begin{array}{cccc} x(n) & x(n+1) & \cdots & x(n+M-1) \end{array} \right]^T, \qquad (3)$$

with $(\cdot)^T$ denoting the transpose. Then, assuming that the phases of the harmonics are independent and uniformly distributed on the interval $(-\pi, \pi]$, the covariance matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$ of the signal in (1) can be written as [11]

$$\begin{aligned} \mathbf{R} &= \mathrm{E}\left\{ \tilde{\mathbf{x}}(n) \tilde{\mathbf{x}}^H(n) \right\} \\ &= \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2 \mathbf{I}, \end{aligned} \qquad (4)$$

*The work of M. G. Christensen is supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–04–0092.

where $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the statistical expectation and the conjugate transpose, respectively. Furthermore, $\mathbf{P}$ is a diagonal matrix containing the squared amplitudes, i.e.,

$$\mathbf{P} = \mathrm{diag}\left(\begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_k \end{bmatrix}\right), \tag{5}$$

with $\mathbf{p}_k = \begin{bmatrix} A_{k,1}^2 & \cdots & A_{k,L_k}^2 \end{bmatrix}$. Defining the total number of sinusoidal components as $Q = \sum_{k=1}^{K} L_k$ and assuming that all the frequencies are distinct, the full rank Vandermonde matrix $\mathbf{Z} \in \mathbb{C}^{M \times Q}$ is defined as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{A}_K \end{bmatrix}, \tag{6}$$

with

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}(\omega_k) & \cdots & \mathbf{a}(\omega_k L_k) \end{bmatrix}, \tag{7}$$

where $\mathbf{a}(\omega) = \begin{bmatrix} 1 & e^{j\omega} & \cdots & e^{j\omega(M-1)} \end{bmatrix}^T$. Also, σ^2 denotes the variance of the additive noise, $w(n)$, and $\mathbf{I}$ is the $M \times M$ identity matrix. We note that $\mathbf{Z}\mathbf{P}\mathbf{Z}^H$ has rank Q. Let

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \tag{8}$$

be the EVD of the $M \times M$ covariance matrix. Then, $\mathbf{U}$ contains the M orthonormal eigenvectors of $\mathbf{R}$, i.e., $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}$ and $\mathbf{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues, λ_k, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. Let $\mathbf{G}_Q$ be formed from the eigenvectors corresponding to the $M - Q$ least significant eigenvalues, i.e.,

$$\mathbf{G}_Q = \begin{bmatrix} \mathbf{u}_{Q+1} & \cdots & \mathbf{u}_M \end{bmatrix}. \tag{9}$$

The noise subspace $\mathbf{G}_Q$ will then be orthogonal to $\mathbf{Z}$, i.e.,

$$\mathbf{Z}^H \mathbf{G}_Q = \mathbf{0}. \tag{10}$$

This is the orthogonality principle of MUSIC and it can be used for finding model parameters and subspace ranks. In practice, this orthogonality will hold only approximately and can be measured using the Frobenius norm.

3. MULTI-PITCH HARMONIC MUSIC

Having introduced the covariance matrix model and the MUSIC orthogonality principle, we now proceed to present the proposed method. Estimates are obtained using MUSIC as the frequencies minimizing the cost function

$$J = \frac{\|\mathbf{Z}^H \mathbf{G}_Q\|_F^2}{MQ(M-Q)}, \tag{11}$$

with $\|\cdot\|_F$ denoting the Frobenius norm and $MQ(M-Q)$ being an order-dependent scaling. If this scaling is omitted, the estimator can easily be observed to be biased. For more on MUSIC and its performance see, e.g., [12, 13]. The set of fundamental frequencies can be found as (see [5])

$$\{\hat{\omega}_k\}_{k=1}^K = \arg \min_{\{\omega_k\}_{k=1}^K} \min_{\{L_k\}_{k=1}^K} \frac{\|\mathbf{Z}^H \mathbf{G}_Q\|_F^2}{MQ(M-Q)} \tag{12}$$

$$= \arg \min_{\{\omega_k\}_{k=1}^K} \min_{\{L_k\}_{k=1}^K} \sum_{k=1}^{K} \frac{\|\mathbf{A}_k^H \mathbf{G}_Q\|_F^2}{MQ(M-Q)}. \tag{13}$$

Since the minimization is carried out over independent, non-negative terms, the optimization problem can be rewritten as

$$\{\hat{\omega}_k\}_{k=1}^K = \arg \min_{\{L_k\}_{k=1}^K} \min_{\{\omega_k\}_{k=1}^K} \sum_{k=1}^{K} \frac{\|\mathbf{A}_k^H \mathbf{G}_Q\|_F^2}{MQ(M-Q)} \tag{14}$$

$$= \arg \min_{\{L_k\}_{k=1}^K} \sum_{k=1}^{K} \min_{\omega_k} \frac{\|\mathbf{A}_k^H \mathbf{G}_Q\|_F^2}{MQ(M-Q)}, \tag{15}$$

meaning that the fundamental frequencies can be found independently. However, it is worth stressing that it is necessary to ensure that $\omega_k \neq \omega_l$, for $k \neq l$. Also note that the set of possible frequencies of ω_k depends on the candidate order L_k. Independent minimizations over the fundamental frequencies is a significant advantage since the set of possible fundamental frequencies is large, while the set of harmonics is small in comparison. The above minimizations can be solved efficiently by treating Q as an independent variable and then calculating

$$S(\omega_k, L_k, Q) = \|\mathbf{A}_k^H \mathbf{G}_Q\|_F^2 \tag{16}$$

for various combinations of $\{\omega_k, L_k, Q\}$ using the FFT-based implementation described in [5]. Note that this optimization otherwise involves a multi-dimensional evaluation of the cost function over all combinations of $\{\omega_k\}_{k=1}^K$ and $\{L_k\}_{k=1}^K$. The number of sources can also be determined using this method for more than one source, i.e., $K > 1$, by allowing $L_k = 0$ for the other sources. For more on order estimation using the MUSIC orthogonality principle and its performance, we refer the interested reader to [5]. For the special case where the number of harmonics $L_k = L$, $\forall k$, is known and equal, i.e., $Q = KL$, the method reduces to the one-dimensional minimization over ω, i.e.,

$$\|\mathbf{A}_k^H \mathbf{G}_Q\|_F^2 = \sum_{l=1}^{L} \mathbf{a}^H(\omega l) \mathbf{G}_Q \mathbf{G}_Q^H \mathbf{a}(\omega l), \tag{17}$$

where the fundamental frequencies can be identified as the K deepest valleys in the cost function. The complexity of the proposed estimator can be reduced by first finding the appropriate ranks of the signal and noise subspaces using unconstrained frequencies. This can be done using a simple special case of the proposed estimator, where first it is assumed that $L_k = 1$, $\forall k$. Then, the total number of sinusoids, and thereby the signal and noise subspace ranks, can be estimated as the number of sources K. The harmonic MUSIC algorithm can then be applied given these subspace ranks, whereby the total number of different combinations of orders $\{L_k\}$ is greatly reduced.

For a given or estimated L_k, the gradient of the cost function (11) can be shown to be (for simplicity we here drop the scaling)

$$\nabla J \triangleq \frac{\partial J}{\partial \omega_0} = 2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{A}_k^H \mathbf{G}_Q \mathbf{G}_Q^H \frac{\partial \mathbf{A}_k}{\partial \omega_0}\right\}\right), \tag{18}$$

with $\mathrm{Re}(\cdot)$ denoting the real value, $\odot$ the Schur-Hadamard product, and

$$\frac{\partial \mathbf{A}_k}{\partial \omega_0} = \mathbf{Y}_k \odot \mathbf{A}_k \tag{19}$$

with

$$\mathbf{Y}_k = \begin{bmatrix} 0 & \cdots & 0 \\ j & \cdots & jL_k \\ \vdots & \vdots & \vdots \\ j(M-1) & \cdots & j(M-1)L_k \end{bmatrix}. \tag{20}$$

Similarly, the Hessian can be derived to be

$$\nabla^2 J \triangleq \frac{\partial^2 J}{\partial \omega_0^2} \tag{21}$$

$$= 2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{A}_k^H \mathbf{G}_Q \mathbf{G}_Q^H \left(\mathbf{Y}_k \odot \mathbf{Y}_k \odot \mathbf{A}_k\right)\right.\right. \tag{22}$$

$$\left.\left. + \left(\mathbf{Y}_k \odot \mathbf{A}_k\right)^H \mathbf{G}_Q \mathbf{G}_Q^H \left(\mathbf{Y}_k \odot \mathbf{A}_k\right)\right\}\right). \tag{23}$$

The gradient and the Hessian can be used for finding refined estimates using standard methods. Here, we iteratively find a refined estimate of the fundamental frequency using Newton's method, i.e.,

$$\hat{\omega}_k^{(i+1)} = \hat{\omega}_k^{(i)} - \delta \frac{\nabla J}{\nabla^2 J}, \tag{24}$$

with i being the iteration index and δ a small, positive constant, which is found using approximate line search. The method is initialized for $i = 0$ using the coarse fundamental frequency estimate and order obtained from (15). As the order L_k and the signal subspace rank Q is kept fixed in the Newton method, only the poles of the Vandermonde matrix $\mathbf{A}_k$ changes in each iteration.

4. NUMERICAL RESULTS

Initially, we illustrate the estimator's ability to estimate multiple fundamental frequencies. In Figure 1, two fundamental frequency tracks are shown as a function of time, i.e., segments, with one being stationary while the other increases linearly, while still being stationary within each segment. Both sets of harmonics consist of 5 sinusoids having uniformly distributed phases that are randomized in each segment and unit amplitudes such that the same performance can be expected for the two sources. In the figure, the fundamental frequencies estimated by the proposed method are shown as circles while the true fundamental frequencies are indicated by a solid line. The number of sinusoids is assumed unknown in this experiment. The segment size used was 200 samples corresponding to 25 ms for a sampling frequency of 8000 Hz, a typical segment size for audio and speech applications. The asymptotic CRLB given in (2) can be seen to be inversely proportional to the noise variance, and generally depend on the pseudo signal-to-noise ratio (PSNR), defined as

$$PSNR_k = 10 \log_{10} \frac{\sum_{l=1}^{L_k} A_{k,l}^2 l^2}{\sigma^2}\ [\mathrm{dB}], \tag{25}$$

which depends on the number of harmonics L_k and the amplitudes $\{A_{k,l}\}$. In this experiment, the PSNR was 40 dB. A covariance matrix of size 100 was used. Note that as the two fundamental frequencies intersect, only one fundamental frequency is found by the estimator as the other is set to 0. We now proceed to evaluate the proposed estimator using Monte Carlo simulations, with 200 realizations for each combination of PSNR and N, using synthetic signals. The performance is measured as the root mean square estimation error (RMSE) defined as

$$RMSE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\omega}_0^{(s)} - \omega_0\right)^2}, \tag{26}$$

with ω_0 and $\hat{\omega}_0^{(s)}$ being the true fundamental frequency and the estimate of the sth Monte Carlo trial. It is compared to the asymptotic CRLB in (2). In each Monte Carlo trial, a signal is generated according to the signal model in (1) with phase and noise realizations being randomized for two sources having well-separated fundamental frequencies of 0.1650 and 0.3937. A covariance matrix of size $M = \lfloor N/2 \rfloor$ was used throughout these experiments. For the sake of simplicity, the orders, $\{L_k\}$ are assumed known in these experiment and is set to 3 for both sources. First, we observe the RMSE as a function of the PSNR, with a fixed number of observations, namely $N = 400$. The results are shown in Figure 2. Then, as depicted in Figure 3, the RMSE is investigated as a function of the number of observations, N, for a fixed PSNR of 40 dB. As can be seen from both figures, the proposed estimator performs well having a variance close to the CRLB. Also, an experiment was conducted to investigate how the estimator performs as the fundamental frequencies of two sources approach each other. Figure 4 shows the RMSE as a function of the difference between the fundamental frequencies, i.e., $\Delta = |\omega_1 - \omega_2|$, for a PSNR of 40 dB and $N = 160$.

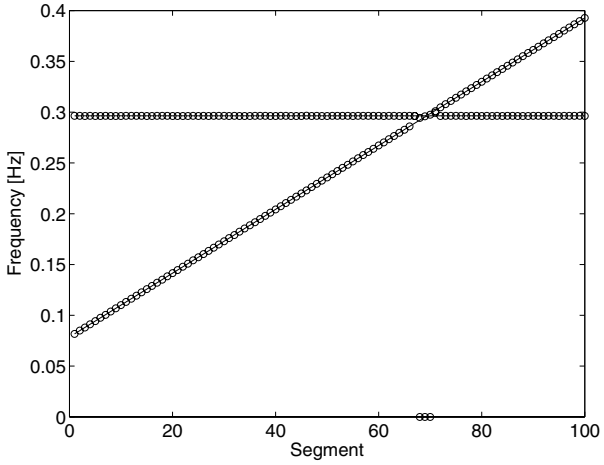

Fig. 1. True (solid) and estimated (circles) fundamental frequencies for two sources having unknown orders.

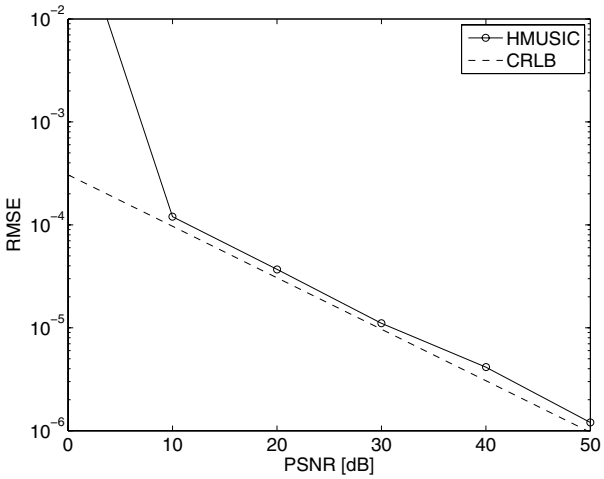

Fig. 2. RMSE and CRLB as a function of PSNR for $N = 400$ and two sources having known orders.

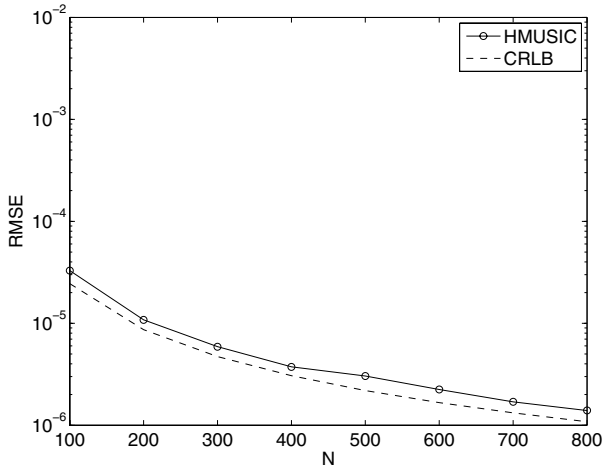

Fig. 3. RMSE and CRLB as a function of the number of observations, N, for $PSNR = 40$ dB and two sources having known orders.

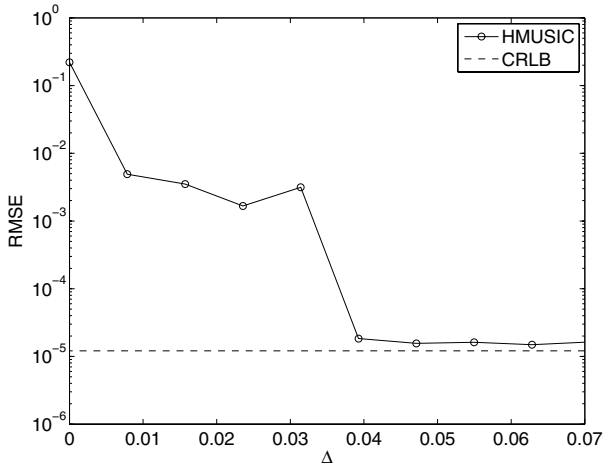

Fig. 4. RMSE and CRLB as a function of the difference, Δ, between the fundamental frequencies of two sources having known orders for $N = 160$ and $PSNR = 40$ dB.

5. CONCLUSION

A method for estimating the fundamental frequencies of multiple sets of periodic waveforms in white Gaussian noise has been proposed. The method, which is based on the MUSIC orthogonality principle, also estimates the order, i.e., the number of harmonics for each member of the set, and can also be used for finding the number of sources. The performance of the estimator has been evaluated using Monte Carlo simulations, and it has been found that the proposed estimator has good statistical performance approaching the Cramér-Rao lower bound.

6. REFERENCES

[1] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(5), pp. 1124–1138, Oct. 1986.

[2] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Processing Lett.*, vol. 11(7), pp. 609–612, July 2004.

[3] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[4] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, "Subspace-based fundamental frequency estimation," in *Proc. European Signal Processing Conf.*, 2004, pp. 637–640.

[5] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. on Audio, Speech and Language Processing*, Apr. 2006, submitted.

[6] R. Gribonval and E. Bacry, "Harmonic Decomposition of Audio Signals with Matching Pursuit," *IEEE Trans. Signal Processing*, vol. 51(1), pp. 101–111, Jan. 2003.

[7] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[8] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 804–816, 2003.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34(3), pp. 276–280, Mar. 1986.

[10] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[11] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[12] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(5), pp. 720–741, May 1989.

[13] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound: Futher results and comparisons," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38(12), pp. 2140–2150, Dec. 1990.